Discrete Choice and Rational Inattention: a General Equivalence Result*

Mogens Fosgerau
Technical University of Denmark

Emerson Melo
Indiana University

Matthew Shum
Caltech

February 3, 2017

Preliminary and incomplete: Comments welcome

Abstract

This paper establishes a general equivalence between discrete choice and rational inattention models. We show that the choice probabilities emerging from any random utility discrete choice model can be obtained from a class of suitably generalized rational inattention models, and vice versa. Thus any discrete choice model can be given an interpretation in terms of boundedly rational behavior. The underlying idea is that the surplus function of a discrete choice model has a convex conjugate that is a generalized entropy (which is a suitable generalization of the Shannon entropy funcion). These generalized entropies are used to construct an information cost function for a generalized rational inattention model. We denote this class of rational inattention problems as *Generalized Entropic Rational Inattention* (GERI) models.

JEL codes: D03, C25, D81, E03

Keywords: Rational Inattention, discrete choices, general entropy, convex analysis

1 Motivation

Rational inattention has become an important paradigm for modelling boundedly rational behavior in many areas of economics (Sims, 2010). In this note we develop a general equivalence between discrete (multinomial) choice and rational inattention models. This

*First draft: December 22, 2016. We thank Marcus Berliant, Mark Dean, and Ryan Webb for useful comments. Alejandro Robinson Cortes provided research assistance.

extends a connection between rational inattention and the multinomial logit model derived by Matejka and McKay (2015). This is important for several reasons. First, we show that the connection with discrete choice models is a typical feature of the rational inattention model, and that the Matejka and McKay (2015) result is a specific example. Second, given the empirical relevance of discrete choice models, it is useful to know that a rational inattention model can generate the same choice probabilities as any particular discrete choice model. Moreover, by exploiting convex analytic properties of the discrete choice model, we show a "duality" between the discrete choice and rational inattention models in the sense of convex conjugacy. This yields new insights into the structure of the rational inattention model. More specifically, we utilize a class of "generalized entropy" functions (Fosgerau and de Palma, 2016) to generalize the information cost function in the rational inattention model in a manner leading to choice probabilities that are consistent with discrete choice models; this connection results in model with tractable choice probabilities, which facilitates their analysis and use in empirical applications. We denote this class of rational inattention problems as *Generalized Entropic Rational Inattention* (GERI) models.[1]

Despite this equivalence, however, random utility discrete choice models and rational inattention models are *not* the same (as pointed out by Matejka and McKay (2015), Caplin, Dean, and Leahy (2016)). We also characterize the full solution of the rational inattention model with our generalized information cost function, and highlight important distinct features of choice under rational inattention, including the possibility that some options may be chosen with zero probabilities, and that a certain "regularity" property that random utility discrete choice models may not hold for a rational inattention model.

In Section 2 we start with the random utility discrete choice model, and introduce the class of generalized entropy functions. Section 3 introduces the rational inattention model. We show how the generalized entropy functions can be used to model the information cost function in the rational inattention model. Section 4 presents the equivalence between choice probabilities emerging from the discrete choice model, and those emerging from the rational inattention model based on the generalized entropy functions. In Section 5 we characterize the full solution of the rational inattention model, and highlight important differences vis-a-vis the discrete choice model. Section 6 concludes.

[1] This complements work by Hébert and Woodford (2016), who also consider generalizations of the information cost function.

2 Discrete choice model

Consider a decision-maker (DM) making discrete choices among a set of $i = 1, \ldots, N$ options, where, for each option i, the utility is given by

$$u_i = \tilde{v}_i + \epsilon_i, \tag{1}$$

where $\tilde{\mathbf{v}} = (\tilde{v}_1, \ldots, \tilde{v}_N)$ is deterministic and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)$ is a vector of random utility shocks. This is the classic additive random utility framework pioneered by McFadden (1978).

Assumption 1. *The random vector $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)$ follows a joint distribution that is absolutely continuous, independent of $\tilde{\mathbf{v}}$, and fully supported on $\mathbb{R}^N$ for $j = 1, \ldots, N$.*

An important concept in this note is the *surplus function* of the discrete choice model (so named by McFadden, 1981), defined as

$$W(\tilde{\mathbf{v}}) = \mathbb{E}_{\boldsymbol{\epsilon}}(\max_i [\tilde{v}_i + \epsilon_i]). \tag{2}$$

As it well-known, $W(\tilde{\mathbf{v}})$ is convex and differentiable. In particular, under assumption 1, the choice probabilities can be expressed as:

$$\frac{\partial W(\tilde{\mathbf{v}})}{\partial \tilde{v}_i} = q_i(\tilde{\mathbf{v}}) \equiv \mathbb{P}\left(\tilde{v}_i + \epsilon_i \geq \tilde{v}_j + \epsilon_j, \, \forall j \neq i\right) \quad \text{for } i = 1, \ldots, N.$$

This is the Daly-Zachary-Williams theorem, famous in the discrete choice literature (McFadden, 1978, 1981).

When the ϵ_i's are distributed i.i.d. across options i according to the type 1 extreme value distribution, then the resulting choice probabilities take the well-known multinomial logit form: $q_i(\tilde{\mathbf{v}}) = e^{\tilde{v}_i} / \sum_j e^{\tilde{v}_j}$. Assumption 1 above leaves the distribution of the ϵ's unspecified, thus allowing for choice probabilities beyond the multinomial logit case.

We begin by defining a vector-valued function $\mathbf{H}(\cdot) = (H_1(\cdot), ..., H_N(\cdot)) : \mathbb{R}^N_+ \to \mathbb{R}^N_+$ as the gradient of the exponentiated surplus, i.e.

$$H_i(e^{\tilde{\mathbf{v}}}) = \nabla_{\tilde{v}_i} \left(e^{W(\tilde{\mathbf{v}})} \right). \tag{3}$$

We next present two results from Fosgerau and de Palma (2016) that provide a general

expression for choice probabilities in discrete choice models, and connect that to a class of *generalized entropies* with corresponding generating functions.

Proposition 1. *The choice probabilities for the discrete choice model take the form*

$$q_i(\tilde{\mathbf{v}}) = \frac{H_i\left(e^{\tilde{\mathbf{v}}}\right)}{\sum_{j=1}^{N} H_j\left(e^{\tilde{\mathbf{v}}}\right)}, \quad \forall i, \tag{4}$$

and the surplus function is

$$W(\tilde{\mathbf{v}}) = \log\left[\sum_{i=1} H_i\left(e^{\tilde{\mathbf{v}}}\right)\right]. \tag{5}$$

The function $\mathbf{H}$ *is globally invertible.*

This proposition shows that the choice probabilities emerging from a random utility discrete choice model have a "logit-like" structure (4), expressed in terms of the function $\mathbf{H}(\cdot)$ that is related to the surplus function $W(\tilde{\mathbf{v}})$.

Fosgerau and de Palma (2016) consider functions defined by

$$\mathbf{S}(\mathbf{q}) = \mathbf{H}^{-1}(\mathbf{q}). \tag{6}$$

Any function S defined in this way has a number of useful properties, summarized in the following proposition. Let Δ denote the unit simplex in $\mathbb{R}^N$.

Proposition 2. *Let assumption 1 hold. Then the vector valued-function* $\mathbf{S}(\mathbf{q})$ *defined by (6) satisfies the following conditions.*

(*i*) $\mathbf{S}$ *is continuous and homogenous of degree 1.*

(*ii*) $\mathbf{q} \cdot \log \mathbf{S}(\mathbf{q})$ *is convex.*

(*iii*) $\mathbf{S}$ *is differentiable at any* $\mathbf{q} \in relint(\Delta)$ *with :*

$$\sum_{i=1}^{N} q_i \frac{\partial \log S_i(\mathbf{q})}{\partial q_k} = \theta, k \in \{1, \ldots, N\}$$

for $\theta > 0$ *a scalar constant invariant across choices* k*.*

(*iv*) $\mathbf{S}$ *is globally invertible.*

(*v*) *The convex (Fenchel) conjugate function for the surplus function* $W(\tilde{\mathbf{v}})$ *is*

$$W^*(\mathbf{q}) = \begin{cases} \mathbf{q} \cdot \log \mathbf{S}(\mathbf{q}) & \mathbf{q} \in \Delta \\ +\infty & \textit{otherwise}, \end{cases}$$

Note that part (*iii*) allows us to write $-\frac{\partial(\mathbf{q} \log \mathbf{S}(\mathbf{q}))}{\partial q_k} = \log S_k(\mathbf{q}) + \theta$, which will be important in the sequel.

Part (*v*) establishes a close relationship between the function $\mathbf{S}(\cdot)$ and the surplus function $W(\tilde{\mathbf{v}})$ of the corresponding discrete choice model; this relationship is in terms of convex conjugacy (Rockafellar, 1970, ch. 12). As a leading example, consider the multinomial logit model. In the multinomial logit model, $\boldsymbol{\epsilon}$ is an i.i.d. extreme value type 1 random vector. Then the surplus $W(\tilde{\mathbf{v}}) = \log\left(\sum_{j=1}^N e^{\tilde{v}_j}\right)$ and the function $\mathbf{S}(\mathbf{q}) = \mathbf{q}$ is just the identity. The convex conjugate of the surplus is $W^*(\mathbf{q}) = \mathbf{q} \cdot \log \mathbf{q}$, which means that $-W^*(\mathbf{q})$ is just the Shannon (1948) entropy.

For a general discrete choice model (1) satisfying assumption 1, $-W^*$ is then a generalized entropy. The corresponding function $\mathbf{S}$ generates the generalized entropy through Proposition 2(v). Furthermore, by Fenchel's equality (cf. Rockafellar, 1970, Thm. 23.5), we also have

$$W(\tilde{\mathbf{v}}) = \mathbf{q} \cdot \tilde{\mathbf{v}} - W^*(\mathbf{q}) \tag{7}$$

for $\nabla W(\tilde{\mathbf{v}}) = \mathbf{q}$. Note that $W(\tilde{\mathbf{v}}) = \sum_{i=1}^N q_i(\tilde{\mathbf{v}})(\tilde{v}_i + \mathbb{E}(\epsilon_i | u_i \geq u_j, j \neq i))$. Combining this latter expression with (7), we obtain an alternative expression for the conjugate function[2]:

$$W^*(\mathbf{q}) = \mathbf{q} \cdot \log \mathbf{S}(\mathbf{q}) = -\sum_i q_i \mathbb{E}[\epsilon_i | u_i \geq u_j, j \neq i].$$

This establishes a connection between the generalized entropy function $\mathbf{S}(\mathbf{q})$ and the joint distribution of $\boldsymbol{\epsilon}$, the random utility shocks.[3]

3 Rational inattention

We now introduce the rational inattention model, as presented in Matejka and McKay (2015) and Hébert and Woodford (2016). The decision maker is again presented with

[2] See, for instance, Chiong, Galichon, and Shum (2016).

[3] Additionally, we conjecture that $\log S_i(\mathbf{q}) = -\mathbb{E}[\epsilon_i | u_i \geq u_j, j \neq i]$ for $i = 1, \ldots, N$, but have not proved it. For the multinomial logit case, corresponding to $\mathbf{S}(\mathbf{q}) = \mathbf{q}$, McFadden (1978) showed that $\gamma - \log q_i = \mathbb{E}[\epsilon_i | u_i \geq u_j, j \neq i]$, for γ being Euler's constant.

a group of N options, from which he must choose one. Each option has an associated payoff $\mathbf{v} = (v_1, ..., v_N)$, but in contrast to the discrete choice model, the vector of payoffs is unobserved by the DM. Instead, the DM considers the payoff vector $\mathbf{V}$ to be random, taking values in $\mathcal{V} \subset \mathbb{R}^N$; for simplicity, we take $\mathcal{V}$ to be finite. The DM possesses some prior knowledge about the available options, given by a probability measure $\mu(\mathbf{v}) = \Pr(\mathbf{V} = \mathbf{v})$.

The DM's choice is represented as a random action $\mathbf{A}$ that is a canonical unit vector in $\mathbb{R}^N$. The payoff resulting from the action is $\mathbf{V} \cdot \mathbf{A}$, namely that value of the entry in $\mathbf{V}$ indicated by the action $\mathbf{A}$.

The problem of the rationally inattentive DM is to choose the conditional distribution $\Pr(\mathbf{A}|\mathbf{V})$, balancing the expected payoff against the cost of information. At the least informative extreme, $\mathbf{A}$ is independent of $\mathbf{V}$ such that $\Pr(\mathbf{A}|\mathbf{V}) = \Pr(\mathbf{A})$. At the other extreme, $\mathbf{A}$ identifies the option with maximal payoff.

Intuitively, the rational inattention problem may be described as follows. The DM is endowed with the prior belief μ about the possible realizations of $\mathbf{V}$. Then the DM receives a signal $\mathbf{s}$ on the state $\mathbf{V}$ to update his prior belief μ. The joint distribution of priors and signals define an information strategy with the property that the marginal distribution over states equals the DM's prior μ, which ensures that the DM's posterior beliefs are consistent with prior. Thus, given these restrictions the DM is only free to choose the conditional distribution of signals. Finally, the DM maximizes expected payoffs, induced by the joint distribution of states and signals, minus the information cost. This cost captures the fact signals may have different degrees of precision.

In general, the presence of signals makes the RI problem a complicated variational problem. Fortunately, we can use the fact that as each action is associated with a particular signal, the cost of information is given by the mutual information between states and actions. This fact allows us to study a RI problem in terms of actions and states.[4]

According to this, denote an action by i and write $p_i(\mathbf{v})$ as shorthand for $\Pr(\mathbf{A} = i|\mathbf{V} = \mathbf{v})$. The DM's strategy is a solution to the following variational problem:

$$\max_{\{p_i(\cdot)\}} \left(\sum_{\mathbf{v} \in \mathcal{V}} \left(\sum_{i=1}^{N} v_i p_i(\mathbf{v}) \right) \mu(\mathbf{v}) - \text{information cost} \right), \tag{8}$$

[4]For further details we refer the reader to Matejka and McKay (2015, pp. 277–280) and Hébert and Woodford (2016, p.10).

subject to

$$p_i(\mathbf{v}) \geq 0 \quad , \tag{9}$$

$$\sum_{i=1}^{N} p_i(\mathbf{v}) = 1 \quad . \tag{10}$$

The previous literature has used the Shannon entropy to specify the information cost. As shown by Matejka and McKay (2015), this connects the rational inattention model to the logit model. We present this connection in the next Section 3.1. Then in Section 3.2 we extend this by introducing generalized entropy to the problem.

3.1 Shannon entropy and multinomial logit

The key element in the program above is the modelling of information processing. Let $\kappa(\mathbf{p}, \mu)$ denote a function that measures the amount of information processed. It will depend on the vector of choice probabilities $\mathbf{p}(\mathbf{v})$ and the prior beliefs μ.

We denote for convenience the Shannon entropy by $\Omega(\mathbf{p}) = -\mathbf{p} \cdot \log \mathbf{p}$. We also use $\mathbf{p}(\mathbf{v}) = (p_1(\mathbf{v}), \ldots, p_N(\mathbf{v}))$ for the vector of choice probabilities conditional on $\mathbf{V} = \mathbf{v}$, while the expected values of the conditional choice probabilities are denoted $p_i^0 = \mathbb{E}p_i(\mathbf{V})$ and $\mathbf{p}^0 = (p_1^0, \ldots, p_N^0)$.

Matejka and McKay (2015) propose to measure the amount of information processed by the mutual (Shannon) information between $\mathbf{V}$ and the actions $\mathbf{A}$. It may be written as

$$\kappa(\mathbf{p}, \mu) = \Omega(\mathbb{E}(\mathbf{p}(\mathbf{V})) - \mathbb{E}(\Omega(\mathbf{p}(\mathbf{V})))) \tag{11}$$

$$= -\sum_{i=1}^{N} p_i^0 \log p_i^0 + \sum_{\mathbf{v} \in \mathcal{V}} \left(\sum_{i=1}^{N} p_i(\mathbf{v}) \log p_i(\mathbf{v}) \right) \mu(\mathbf{v}). \tag{12}$$

Matejka and McKay then specify the information cost as $\lambda\kappa(\mathbf{p}, \mu)$ where $\lambda > 0$ is the unit cost of information. The choice strategy of the rationally inattentive DM is the collection of conditional probabilities $\mathbf{p} = \{p_i(\mathbf{v})\}_{i=1}^N$ that solves the optimization problem

$$\max_{\{p_i(\mathbf{v})\}_{i=1}^N} \sum_{\mathbf{v} \in \mathcal{V}} \left(\sum_{i=1}^{N} v_i p_i(\mathbf{v}) \right) \mu(\mathbf{v}) - \lambda\kappa(\mathbf{p}, \mu) \tag{13}$$

subject to (9) and (10). The DM solves (16) to find conditional choice probabilities

$$p_i(\mathbf{v}) = \frac{p_i^0 e^{v_i/\lambda}}{\sum_{j=1}^N p_j^0 e^{v_j/\lambda}} \quad \text{for } i = 1, \ldots, N. \tag{14}$$

When the prior μ is constant then also the p_i^0 are constant. In this case the rational inattention model reduces to the multinomial logit model. This is a key observation from Matejka and McKay (2015).

3.2 Moving beyond Shannon: the *Generalized Entropic Rational Inattention* (GERI) models

The goal of this paper is to generalize Matejka and McKay's equivalence result between rational inattention and multinomial logit. To achieve that, we replace the Shannon entropy by the generalized entropy introduced in Section 2 above. This is reasonable for two reasons. First, generalized entropy leads to information costs with desirable properties. Second, since each generalized entropy implies a corresponding discrete choice model (Proposition 2), it turns out that each RI model with an information cost derived from a generalized entropy will generate choice probabilities consistent with a corresponding discrete choice model (Proposition 4 below); this implies that *any* discrete choice model can be microfounded by a corresponding rational inattention model, thus generalizing substantially a main result in Matejka and McKay (2015). We discuss each point in turn.

We begin by generalizing the rational inattention framework described above, using generalized entropy in place of Shannon entropy. Specifically, we let $\mathbf{S}$ be the entropy generator corresponding to some discrete choice model satisfying Assumption 1 and define $\Omega_{\mathbf{S}}(\mathbf{p}) = -\mathbf{p} \cdot \log \mathbf{S}(\mathbf{p})$ as the corresponding generalized entropy. Accordingly, we define a general information cost by

$$\begin{aligned} \kappa_{\mathbf{S}}(\mathbf{p}, \mu) &= \Omega_{\mathbf{S}}\left(\mathbf{p}^{\mathbf{0}}\right) - \mathbb{E}_\mu \Omega_{\mathbf{S}}(\mathbf{p}(\mathbf{V})) \\ &= -\mathbf{p}^0 \cdot \log \mathbf{S}\left(\mathbf{p}^0\right) + \sum_{\mathbf{v} \in \mathcal{V}} \left[\mathbf{p}(\mathbf{v}) \cdot \log \mathbf{S}(\mathbf{p}(\mathbf{v}))\right] \mu(\mathbf{v}). \end{aligned} \tag{15}$$

Accordingly, we consider a generalized class of rational inattention models in which the DM

chooses the collection of conditional probabilities $\mathbf{p} = \{p_i(\mathbf{v})\}_{i=1}^N$ to optimize

$$\max_{\{p_i(\mathbf{v})\}_{i=1}^N} \sum_{\mathbf{v}\in\mathcal{V}} \left(\sum_{i=1}^N v_i p_i(\mathbf{v}) \right) \mu(\mathbf{v}) - \lambda \kappa_S(\mathbf{p}, \mu) \tag{16}$$

where $\mathbf{S}$ corresponds to a discrete-choice model; that is, the function $\mathbf{S}$ is defined as $\mathbf{S} = \mathbf{H}^{-1}$ for some $\mathbf{H}$ function satisfying Proposition 1. We refer to this class as ***Generalized Entropic Rational Inattention* (GERI)** models.

Before characterizing the solutions of GERI models, we describe some features of the generalized entropic information cost function.

3.2.1 Properties of $\kappa_{\mathbf{S}}(\mathbf{p}, \mu)$

The general information cost $\kappa_{\mathbf{S}}(\mathbf{p}, \mu)$, as defined in Eq. (15), possesses certain reasonable and desirable properties that have been discussed in the existing literature. First, the action $\mathbf{A}$ carries no information about the payoff $\mathbf{V}$ when $\mathbf{A}$ and $\mathbf{V}$ are independent. In that case the information cost should be zero.

Condition 1. ***Independence***. *If* $\mathbf{A}$ *and* $\mathbf{V}$ *are independent, then* $\kappa_{\mathbf{S}}(\mathbf{p}, \mu) = 0$.

Secondly, the mutual Shannon information $\kappa(\mathbf{p}, \mu)$ is a convex function of $\mathbf{p}$. This is useful as it ensures a unique solution to the problem of the rationally inattentive DM. We will show that the information cost $\kappa_{\mathbf{S}}(\mathbf{p}, \mu)$ has a slightly weaker property, namely that it is convex on sets where $\mathbf{p}^0$ is constant (precisely, sets of mean-preserving choice probability vectors).

Condition 2. ***Convexity***. *For a given* μ*, the information cost function* $\kappa_S(\mathbf{p}, \mu)$ *is convex on any set of choice probabilities vectors satisfying* $\left\{\mathbf{p} : \mathcal{V} \to \Delta^N | \ \mathbb{E}_\mu \mathbf{p}(\mathbf{V}) = \mathbf{p}^0\right\}$.

The mutual Shannon information $\kappa(\mathbf{p}, \mu)$ satisfies these two properties. The next proposition establishes that the information cost defined in (15) using the generalized entropy functions also satisfies these properties.

Proposition 3. *The information cost defined in Eq. (15) satisfies conditions 1 and 2.*

This proposition shows that the generalized information cost (15) retains some desirable properties from the Shannon-based information cost. The generalized entropy information

cost function may, however, not satisfy all the properties of information cost functions discussed in Hébert and Woodford (2016).[5] At the same time, our use of the generalized entropy information cost functions exploits the close connection between generalized entropy functions and discrete choice models, which allows us to generalize the types of choice probabilities which can arise in rational inattention models far beyond the multinomial logit. This is the topic of the remainder of the paper.

3.2.2 Choice probabilities in GERI models

Next, we characterize the choice probabilities for a GERI model.

Proposition 4. *Consider a GERI problem in which the DM solves the program*

$$\max_{\{p_i(\mathbf{v})\}} \left(\sum_{\mathbf{v}\in\mathcal{V}} \left(\sum_{i=1}^{N} v_i p_i(\mathbf{v}) \right) \mu(\mathbf{v}) - \lambda \kappa_{\mathbf{S}}(\mathbf{p}, \mu) \right), \tag{17}$$

s.t. (9) and (10). The cost function $\kappa_{\mathbf{S}}(\mathbf{p}, \mu)$ is given by

$$\kappa_{\mathbf{S}}(\mathbf{p}, \mu) = -\sum_{i=1}^{N} p_i^0 \log S_i(\mathbf{p^0}) + \sum_{\mathbf{v}\in\mathcal{V}} \left(\sum_{i=1}^{N} p_i(\mathbf{v}) \log S_i(\mathbf{p}(\mathbf{v})) \right) \mu(\mathbf{v}) \tag{18}$$

for a generalized entropy function $\mathbf{S}$.

For the set of choices $\mathcal{I} \subseteq \{1, \ldots, N\}$ such that $p_i^0 > 0$ (nonzero prior probabilities), the conditional choice probabilities satisfy

$$p_i(\mathbf{v}) = \frac{H_i(e^{(\mathbf{v}+\log \mathbf{S}(\mathbf{p}^0))/\lambda})}{\sum_{j\in\mathcal{I}} H_j(e^{(\mathbf{v}+\log \mathbf{S}(\mathbf{p}^0))/\lambda})} \quad \text{for } i \in \mathcal{I}, \tag{19}$$

where $\mathbf{H} = \mathbf{S}^{-1}$.

Proposition 4 generalizes results in Matejka and McKay (2015); compare Eq. (19) to Eq. (14). In fact, when $\mathbf{S}$ is the identity, Eq. (19) reduces to Eq. (14).

[5] In particular, we have only been able to prove the "Blackwell dominance" condition (cf. Hébert and Woodford (2016, pg. 18)) for the generalized entropy information cost functions only under more restrictive conditions.

4 Equivalence between discrete choice and rational inattention

In this section we establish the equivalence between discrete choice models and rational inattention models. In particular, we show that the choice probabilities generated by GERI model lead to the same choice probabilities as a corresponding choice probabilities from a discrete choice model. For convenience, we also assume in the remainder of this paper that $\lambda = 1$. This simplifies notation, and results for $\lambda \neq 1$ can be derived in a straightforward manner.

In order to gain some intuition, return to Eq. (1) and define, for each i, the *perturbed valuation*

$$\tilde{v}_i = v_i + \log S_i(\mathbf{p^0}) \quad \text{for } i = 1, \ldots, N. \tag{20}$$

It is easy to see that plugging in these valuations in Eq. (3), we obtain the choice probabilities

$$q_i(\tilde{\mathbf{v}}) = \frac{H_i\left(e^{\tilde{\mathbf{v}}}\right)}{\sum_{j=1}^{N} H_j\left(e^{\tilde{\mathbf{v}}}\right)}, \quad \forall i. \tag{21}$$

This expression is identical to Eq. (19), which is the solution to the GERI problem (17).

Similarly, solving the rational inattention problem (17) we obtain the choice probabilities (19), which correspond to a discrete choice model with random utilities given by $\tilde{v}_i = v_i + \log S_i(\mathbf{p}^0)$ for $i = 1, \ldots, N$.

Intuitively we have the following

$$p_i(\mathbf{v}) = \frac{H_i(e^{\mathbf{v}+\log \mathbf{S}(\mathbf{p^0})})}{\sum_{j=1}^{N} H_j(e^{\mathbf{v}+\log \mathbf{S}(\mathbf{p^0})})} = \frac{H_i(e^{\tilde{\mathbf{v}}})}{\sum_{j=1}^{N} H_j(e^{\tilde{\mathbf{v}}})} = q_i(\tilde{\mathbf{v}}).$$

It is worth remarking that the priors p_i^0s must be consistent with the probability measure $\mu(\mathbf{v})$:

$$\mathbf{p}^0 = \sum_{\mathbf{v} \in \mathcal{V}} \mathbf{p}(\mathbf{v} + \log \mathbf{S}(\mathbf{p}^0))\mu(\mathbf{v}). \tag{22}$$

Proposition 5. *Every GERI model yielding strictly positive conditional and prior choice probabilities for all options $i = 1, \ldots, N$ is observationally equivalent to (has identical conditional choice probabilities as) a random utility discrete choice model defined by Eqs. (3),(6), and (20) satisfying assumption 1. In particular, for each value $\mathbf{v} \in \mathcal{V}$ we have the following:*

$$p_i(\mathbf{v}) = q_i(\tilde{\mathbf{v}}) \quad \textit{for } i = 1, \ldots, N.$$

Proposition 5 establishes that every random utility discrete choice model (1) with surplus function W is observationally equivalent (i.e. has the same choice probabilities) to a GERI model with a function $\mathbf{S}$ defined by Eqs. (3) and (6).

There is an important caveat to this result, as it holds only for GERI models in which all the options are chosen with non-zero probabilities. For RI models based on Shannon entropy, Matejka and McKay (2015) and Caplin, Dean, and Leahy (2016) discuss how the optimal solution is characterized by *consideration sets*; that is, the DM will optimally set the prior choice probabilities on some items to zero so that for some i, $p_i^0 = 0$ and hence $p_i(\mathbf{v}) = 0$ for all $\mathbf{v} \in \mathcal{V}$. Such zero choice probabilities for certain options represents a point of departure between discrete choice and rational inattention models; indeed, as stated in Proposition 5, the equivalence between discrete choice and rational inattention holds only for the set of choices with strictly positive prior choice probabilities. We turn to this point next.

5 Solving the GERI model

The previous section has discussed an equivalence between discrete-choice models, and GERI models, in terms of the choice probabilities which they generate, for the choices which are chosen with positive probability. However, an important distinction between discrete choice and GERI models is that in the latter, some options can be chosen with zero probability, for which the proposition in the previous section does not apply.

In this section, we highlight this feature of the rational inattention model, by considering the complete solution to the rational inattention model. We begin by using convex-analytic tools to derive an alternative characterization of the rational inattention problem, and then characterize some features of the optimizing solution.

Proposition 6. *Let assumption 1 hold and let* $\mathbf{p}(\mathbf{v})$ *be a solution to the GERI problem. Then the following statements hold:*

i) The generalized inattention cost function may be written as

$$\kappa_{\mathbf{S}}(\mathbf{p}, \mu) = W^*(\mathbf{p^0}) - \sum_{\mathbf{v} \in \mathcal{V}} W^*(\mathbf{p}(\mathbf{v}))\mu(\mathbf{v}). \tag{23}$$

ii) The choice probabilities in the GERI model $\mathbf{p}(\mathbf{v})$ *can also be generated, for each* $\mathbf{v} \in \mathcal{V}$*,*

by the problem

$$\max_{\mathbf{p}\in\Delta} \left\{ \mathbf{p} \cdot (\mathbf{v} + \log \mathbf{S}(\mathbf{p}^0)) - W^*(\mathbf{p}) \right\}. \tag{24}$$

Eq. (24) provides an alternative pointwise representation of the GERI problem (8).

iii) The optimal value of the GERI program (17) is equal to

$$\sum_{\mathbf{v}\in\mathcal{V}} W(\mathbf{v} + \log \mathbf{S}(\mathbf{p}^0))\mu(\mathbf{v}). \tag{25}$$

This proposition characterizes the GERI problem in terms of the convex conjugate properties of W, the surplus function from the discrete choice model. It also describes a "duality" between the discrete choice and GERI models, in the sense that the surplus function for the discrete choice model and the generalized information cost (15) are convex conjugates to each other.

The complete solution of the rational inattention model in the Shannon entropy/multinomial logit case, is summarized in Matejka and McKay (2015) and Caplin, Dean, and Leahy (2016). The analogous procedure for solving the GERI model is provided in Proposition 6. Specifically we can solve the full rational inattention problem by optimizing Eq. (25), which is

$$\sum_{\mathbf{v}\in\mathcal{V}} \mu(\mathbf{v}) W\left(\mathbf{v} + \log \mathbf{S}\left(\mathbf{p}^0\right)\right) = \sum_{\mathbf{v}\in\mathcal{V}} \mu(\mathbf{v}) \log \sum_{i=1}^{N} H_i(\exp\left(\mathbf{v} + \log \mathbf{S}(\mathbf{p}^0)\right))$$

(the RHS follows from Eq. (5)). We maximize this subject to $\sum_{i=1}^{N} p_i^0 = 1$ and $p_i^0 \geq 0$ for $i = 1, \ldots, N$, and use the mathematical convention that $0 \log 0 = 0$. Then once $\mathbf{p}^0$ is known, the corresponding choice probabilities are given by Eq. (19).

5.1 Optimal consideration sets

We now consider some features of the prior probabilities emerging from the solution to the full GERI model, as described in the previous section. We begin noticing that the associated Lagrangean may be written as

$$\mathcal{L}(\gamma, \boldsymbol{\xi}) = \sum_{\mathbf{v}} W(\mathbf{v} + \log \mathbf{S}(\mathbf{p}^0))\mu(\mathbf{v}) - \gamma \left[\sum_{i=1}^{N} p_j^0 - 1 \right] + \sum_{i=1}^{N} \xi_j p_j^0,$$

where γ and $\boldsymbol{\xi} = \{\xi_i\}_{i=1}^N$ are the associate Lagrange multipliers.

Then taking first order conditions we get

$$\sum_{\mathbf{v}} \sum_{j=1}^{N} p_j(\mathbf{v}) \frac{\partial \log S_j(\mathbf{p}^0)}{\partial p_i^0} + \gamma + \xi_i = 0 \quad \text{for } i = 1, \ldots, N.$$

Substituting $\sum_{\mathbf{v}} p_j(\mathbf{v})\mu(\mathbf{v}) = \mathbb{E}_\mu p_j(\mathbf{v}) = p_j^0$, we have

$$\sum_{j=1}^{N} p_j^0 \frac{\partial \log S_j(\mathbf{p}^0)}{\partial p_i^0} + \gamma + \xi_i = 0 \quad \text{for } i = 1, \ldots, N.$$

For the case of an interior solution $p_i^0 > 0$, we get $\xi_i = 0$ implying that previous condition becomes

$$\sum_{j=1}^{N} p_j^0 \frac{\partial \log S_j(\mathbf{p}^0)}{\partial p_i^0} = -\gamma.$$

Using Proposition 2(iii), it follows:

$$\theta = -\gamma$$

Then we can write down necessary and sufficient conditions for the optimal $\mathbf{p}^0$. In particular, for all i, we have

$$\sum_{j=1}^{N} p_j^0 \frac{\partial \log S_j(\mathbf{p}^0)}{\partial p_i^0} \leq \theta,$$

with equality if $p_i^0 > 0$.

For the case of the logit and nested logit models we have $\theta = 1$, implying that the necessary and sufficient conditions for the optimum can be expressed as:[6]

$$\sum_{j=1}^{N} p_j^0 \frac{\partial \log S_j(\mathbf{p}^0)}{\partial p_i^0} \leq 1.$$

By solving the full problem, the optimizing priors may involve zero probabilities for some of the choices. Obviously, for these choices, the corresponding choice probabilities will also

[6] Noting that the logit model is the particular case of $\mathbf{S}(\mathbf{p}^0) = \mathbf{p}^0$, these conditions generalize the result in Caplin, Dean, and Leahy (2016, Prop.1).

be zero for all $\mathbf{v}$. In this way, the rational inattention framework can generate violations of the "regularity" property which otherwise characterizes additive random utility discrete choice models (see Example 3 below).[7]

While the optimal consideration sets emerging from the full solution to this problem is difficult to characterize, the following proposition describes one important feature of optimal consideration sets: that they will exclude choices which offer the lowest utility in all states of the world.

Proposition 7. *For some option a, and for all $\mathbf{v} \in \mathcal{V}$, let $v_a \leq v_i$ for all $i \neq a$. Assume that the valuations are all distinct with positive probability. Then $p_a^0 = 0$ (that is, option a is not in the optimal consideration set).*

5.2 Examples

Example 1: Multinomial logit. This is the Matejka and McKay example. For a set of valuations $\tilde{\mathbf{v}}$, the multinomial logit choice probabilities are given by:

$$p_i(\tilde{\mathbf{v}}) = \frac{e^{\tilde{v}_i}}{\sum_{j=1}^{N} e^{\tilde{v}_j}} \tag{26}$$

which arises from a discrete choice model in which the utility shocks ϵ are distributed i.i.d. according to the Type 1 extreme value distribution.

Using the results above, these choice probabilities (26) are also equivalent to those from a rational inattention model with the Shannon entropy function, which is the case $S_i(\mathbf{p}) = p_i$, for each i, and corresponding valuations

$$v_i = \tilde{v}_i - \log p_i^0, \quad i \in \{1, \ldots, n\}.$$

Caplin, Leahy, and Matejka (2016) use this equivalence for the estimation of rational inattention models with Shannon entropy.

Moreover, for this case, we have that the surplus function

$$W(\tilde{\mathbf{v}}) = \log \sum_i \exp(\tilde{v}_i) = \log \left[\sum_i \exp(v_i) * p_i^0 \right].$$

[7] See also Matejka and McKay (2015, pp. 293ff). The regularity property is that adding an option to a choice set cannot increase the choice probability for any of the original choices.

Thus for this case, the alternative representation of the rational inattention program given in Proposition 5 (Eq. (25)) is:

$$\sum_{\mathbf{v}\in\mathcal{V}} \log\left[\sum_i \exp(v_i) * p_i^0\right] \mu(\mathbf{v})$$

which is analogous to maximization problems derived in Matejka and McKay (2015, Eq. (14)) and Caplin, Dean, and Leahy (2016, Eq. (3)).

■

Example 2: Nested logit. From an applied point of view, an important implication of proposition 5 is that it allows us to model complex choice patterns, beyond the multinomial logit case. In this example, we generate nested logit choice probabilities from a GERI model. Among applied researchers, the nested logit model of multinomial choice is preferred over the multinomial logit model because it generates reasonable substitution patterns across products (and avoids the "red bus/blue bus" pitfall).

Following Fosgerau and de Palma (2016), we partition the set of alternatives $i \in \{1,\ldots,N\}$ into mutually exclusive nests, and let g_i denote the nest containing alternative i. Let $\zeta_{g_i} \in (0,1]$ be nest-specific parameters. For a set of valuations $\tilde{\mathbf{v}}$, the nested logit choice probabilities are given by:

$$p_i(\tilde{\mathbf{v}}) = \frac{e^{\tilde{v}_i/\zeta_{g_i}}}{\sum_{j\in g_i} e^{\tilde{v}_j/\zeta_{g_i}}} \cdot \frac{e^{\zeta_{g_i}\log\left(\sum_{j\in g_i} e^{\tilde{v}_j/\zeta_{g_i}}\right)}}{\sum_{\text{all nests } g} e^{\zeta_g \log\left(\sum_{j\in g} e^{\tilde{v}_j/\zeta_g}\right)}}. \tag{27}$$

As is well-known, the nested logit choice probabilities are consistent with a discrete choice model in which the utility shocks ϵ are jointly distributed according to a generalized extreme value distribution.

Using the results above, the nested logit choice probabilities (27) are also equivalent to those from a GERI model with

$$S_i(\mathbf{p}) = p_i^{\zeta_{g_i}} \left(\sum_{j\in g_i} p_i\right)^{1-\zeta_{g_i}} \tag{28}$$

and valuations

$$v_i = \tilde{v}_i - \log S_i(\mathbf{p^0}), \quad i \in \{1,\ldots,n\}.$$

For this case we have that the surplus function $W(\tilde{\mathbf{v}})$ takes the form:

$$W(\tilde{\mathbf{v}}) = \log\left[\sum_g e^{I_g(\tilde{\mathbf{v}})/\zeta_g}\right],$$

where $I_g(\mathbf{v}) = \log\left(\sum_{j\in g} e^{v_j/\zeta_g}\right)$. Thus for this case, the alternative representation of the GERI program given in Proposition 5 (Eq. (25)) is:

$$\sum_{\mathbf{v}\in\mathcal{V}} \log\left[\sum_g e^{I_g(\mathbf{v}+\mathbf{S}(\mathbf{p}^0))/\zeta_g}\right]\mu(\mathbf{v}).$$

We use the previous closed form expression to find the priors $\mathbf{p}^0$. In particular, we solve the following program:

$$\mathcal{L} = \sum_{\mathbf{v}\in\mathcal{V}} \log\left[\sum_g e^{I_g/\zeta_g}\right]\mu(\mathbf{v}) - \gamma\left[\sum_{j=1}^N p_j^0 - 1\right] + \sum_{j=1}^N \xi_j p_j^0.$$

Applying the necessary and sufficient conditions we find that an optimal solution $\mathbf{p}^0$ is characterized by the set of equations:

$$\sum_{j=1}^N p_j^0(\mathbf{v})\frac{\partial \log S_j(\mathbf{p}^0)}{\partial p_i^0} \leq 1 \quad \text{for } i = 1,\ldots,N,$$

with equality if $p_i(\mathbf{v}) > 0$.

■

Example 3: zero prior probabilities and failure of regularity.

Next, we consider a fully solved out example illustrating the possibility of zero prior choice probabilities and failure of regularity, which are results which can happen in the GERI framework but not in the discrete choice model, and represent an important point of difference between the two models. Consider a setting with four choices. Table 5.2 lists the valuation vectors for these four choices in the three equipossible states of the world. We consider both the multinomial logit and nested logit models. (For the nested logit model, we assume that nest 1 consists of choices (2,3) with nesting parameter $\zeta_1 = 0.7$, and nest 2 consists of choices (1,4) with parameter $\zeta_2 = 0.8$.

State:	$\mathbf{v}^1$	$\mathbf{v}^2$	$\mathbf{v}^3$
Choice 1	3	1	3
Choice 2	2	3	3
Choice 3	1	2	2
Choice 4	2	4	2

Table 1: Valuation vectors in Example 3

Model:	MN Logit	MN Logit	Nested Logit	Nested Logit
Choice set:	$\{1,2,3\}$	$\{1,2,3,4\}$	$\{1,2,3\}$	$\{1,2,3,4\}$
p_1^0	0.29	0.51	0.29	0.57
p_2^0	0.71	0.00	0.71	0.00
p_3^0	0.00	0.00	0.00	0.00
p_4^0	—	0.49	—	0.43
Optimized surplus: $\mathbb{E}_{\mathbf{v}} W(\mathbf{v} + \log \mathbf{S}(\mathbf{p}^0))$	2.705	2.865	4.222	6.032

Table 2: Optimal prior probabilities for Example 3

For each model, we compute the optimal prior probabilities first for the choice set $\{1,2,3\}$, and for the expanded choice set $\{1,2,3,4\}$. This example will illustrate how adding choice 4 to the choice set can results in increases in the choice probabilities of choices (1,2,3) thus showing a failure of the regularity property. The optimal prior probabilities are shown in Table 5.2. Qualitatively the results are the same between the logit and nested logit specifications. With the smaller choice set, we see that only choices 1,2 and chosen with positive probability, essentially leading to a consideration set consisting of only those two choices. When choice 4 is added, however, then choice 2 drops out of the consideration set, and only choice 1,4 are chosen with positive probability. This demonstrates a failure of regularity, as the addition of choice 4 *increases* the prior choice probability for choice 1. (Moreover, note that with the expanded choice set, choice 2 is chosen with zero probability, even though it is not inferior in all states of the world, which demonstrates that the characterization of consideration sets in Proposition 7 is not exhaustive.)

Basically, the addition of choice 4 allows agents to form an effective "hedge" in conjunction with choice 1. In the state when choice 1 yields a low payoff (state $\mathbf{v}^2$), choice 4 yields a high payoff; on the contrary, when choice 4 yields a lower payoffs (states $\mathbf{v}^1$ and $\mathbf{v}^3$), choice 1 yields high payoffs.

6 Summary and conclusions

In this paper, we establish a general equivalence between discrete choice and a class of generalized rational inattention models. In particular, we show that the choice probabilities emerging from a random utility discrete choice model can be obtained from a suitable model in the class of Generalized Entropic Rational Inattention (GERI) models, and vice versa. Thus any discrete choice model can be given an interpretation in terms of boundedly rational behavior. The underlying idea is that, by exploiting convex analytic properties of the discrete choice model, we show a "duality" between the discrete choice and GERI models in the sense of convex conjugacy. Precisely, the surplus function of a discrete choice model has a convex conjugate that is a generalized entropy. Thus, GERI models are rational inattention problems in which the information cost functions are constructed from the convex conjugate functions of discrete-choice models.

A few remarks are in order. First, one difference between rational inattention and discrete choice models is that some options may be chosen with zero prior probability in the rational inattention model, which allows for the possibility that the choice probabilities for some existing choices may increase upon introduction of an additional good, which violates the regularity property of random utility models.

Second, there is also a connection between the results here and papers in the decision theory literature. The GERI optimization problem (17) bears resemblance to the variational preferences which Maccheroni, Marinacci, and Rustichini (2006) propose to represent ambiguity averse preferences, as well as to the revealed perturbed utility paradigm proposed by Fudenberg, Iijima, and Strzalecki (2015) to model stochastic choice behavior. Specifically, the information cost function in the rational inattention model appears analogous to the ambiguity aversion indices in the variational preferences, and to the perturbation function in the perturbed utility representations. The main point in this paper is to establish a duality between rational inattention models and random utility discrete choice models, which results in observational equivalence of their choice probabilities, and it seems a similar duality might arise between discrete choice models and these other models from decision theory.

References

A. Caplin, M. Dean, and J. Leahy (2016). Rational Inattention, Optimal consideration sets and stochastic choice. Working paper.

A. Caplin, J. Leahy, and F. Matejka (2016). Rational Inattention and Inference from market Share Data. Working paper.

K. Chiong, A. Galichon, and M. Shum (2016). Duality in Dynamic Discrete Choice Models. *Quantitative Economics*, 7 (1), pp. 83-115.

H. de Oliveira, T. Denti, M. Mihm, K. Ozbek (2015). Rationally Inattentive Preferences and Hidden Information Costs. Working paper.

D. Fudenberg, R. Iijima, and T. Strzalecki (2015). Stochasic Choice and Revealed Perturbed Utility. *Econometrica*, 83 (6), pp. 2371-2409.

M. Fosgerau and A. de Palma (2016). Generalized entropy models. *MPRA Paper No. 70249*.

B. Hébert and M. Woodford (2016). Rational Inattention with Sequential Information Sampling. Working paper.

T. Rockafellar (1970). *Convex Analysis*. Princeton University Press, 1970.

F. Maccheroni, M. Marinacci, and A. Rustichini (2006). Ambiguity Aversion, Robustness, and the Variational Representation of Preferences, *Econometrica*, 74(6): 1447–1498.

F. Matêjka and A. McKay (2015). Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model, *American Economic Review*, 105(1): 272–98.

D. McFadden (1978). Modelling the choice of residential location. In *Spatial Interaction Theory and Residential Location* (A. Karlquist et. al., eds.), North-Holland, Amsterdam.

D. McFadden (1981). Econometric Models of Probabilistic Choice. In: C.Manski and D. McFadden (Eds), Structural Analysis of Discrete Data with Economic Applications, Cambridge, MA: MIT Press, 198–272.

C.E. Shannon (1948). A Mathematical Theory of Communication. *Bell System Technical Journal* 27(3): 379–423.

C. Sims (2010). Rational inattention and monetary economics. *Handbook of Monetary Economics*, Volume 3, pp. 155-181.

Appendix:

A Proofs

Proof of Proposition 1: See Fosgerau and de Palma (2016). □

Proof of Proposition 2: See Fosgerau and de Palma (2016). □

Proof of Proposition 3: *Independence:* By independence, we have, for all i, $p_i(\mathbf{v}) = k_i$, a constant. Then $p_i^0 = k_i$ and $\kappa(\mathbf{p}, \mu) = 0$.

Convexity: We want to show that Ω is convex on any set of actions that has constant $\mathbf{p}^0$. To prove this, consider two sets of choice probabilities $\mathbf{p}_1(\mathbf{v}), \mathbf{p}_2(\mathbf{v})$ where both have the same implied prior probabilities $\mathbf{p}^0$. For $\rho \in [0,1]$, define $\mathbf{p}_\rho$ as the convexification $\rho\mathbf{p}_1(\mathbf{v}) + (1-\rho)\mathbf{p}_2(\mathbf{v})$. Then we would like to show that

$$\rho\kappa(\mathbf{p_1}, \mu) + (1-\rho)\kappa(\mathbf{p_2}, \mu) \geq \kappa(\mathbf{p}_\rho, \mu).$$

But

$$\begin{aligned} &\rho\kappa(\mathbf{p_1}, \mu) + (1-\rho)\kappa(\mathbf{p_2}\,\mu) - \kappa(\mathbf{p}_\rho, \mu) \\ =& -\rho\Omega(\mathbf{p}_1) - (1-\rho)\Omega(\mathbf{p}_2) + \Omega(\rho\mathbf{p}_1 + (1-\rho)\mathbf{p}_1), \end{aligned}$$

which is positive by concavity of $\Omega(\mathbf{p})$ (cf. Proposition 2(ii)).

□

Proof of Proposition 4. The program (17) may be written as

$$\begin{aligned} \mathcal{L}(\mathbf{p}, \boldsymbol{\xi}, \boldsymbol{\mu}) &= \sum_{\mathbf{v}\in\mathcal{V}} \left(\sum_{k=1}^{N} v_k p_k(\mathbf{v}) \right) \mu(\mathbf{v}) - \lambda \left(-\sum_{k=1}^{N} p_k^0 \log S_k(\mathbf{p}^0) + \sum_{\mathbf{v}\in\mathcal{V}} \left(\sum_{k=1}^{N} p_k(\mathbf{v}) \log S_k(\mathbf{p}(\mathbf{v})) \right) \mu(\mathbf{v}) \right) \\ &+ \sum_{\mathbf{v}\in\mathcal{V}} \left(\sum_{k=1}^{n} \boldsymbol{\xi}_k(\mathbf{v}) p_k(\mathbf{v}) \right) \mu(\mathbf{v}) - \sum_{\mathbf{v}\in\mathcal{V}} \boldsymbol{\gamma}(\mathbf{v}) \left[\sum_{k=1}^{n} p_k(\mathbf{v}) - 1 \right] \mu(\mathbf{v}). \end{aligned}$$

Recalling that $p_i^0 = \sum_{\mathbf{v}\in\mathcal{V}} p_i(\mathbf{v})\mu(\mathbf{v})$ for $i = 1, \ldots, n$, the point-wise first-order condition w.r.t $p_i(\mathbf{v})$ can be written as:

$$v_i - \lambda \left(-\log(S_i(\mathbf{p}^0)) - \sum_{k=1}^{n} p_k^0 \frac{\partial \log(S_k(\mathbf{p}^0))}{\partial p_i^0} + \log(S_i(\mathbf{p}(\mathbf{v})) + \sum_{k=1}^{n} p_k(\mathbf{v}) \frac{\partial \log(S_k(\mathbf{p}(\mathbf{v})))}{\partial p_i(\mathbf{v})} \right) + \boldsymbol{\xi}_i(\mathbf{v}) - \boldsymbol{\gamma}(\mathbf{v}) = 0.$$

Under condition **C3** it follows that

$$-\sum_{k=1}^{n} p_k^0 \frac{\partial \log(S_k(\mathbf{p}^0))}{\partial p_i^0} + \sum_{k=1}^{n} p_k(\mathbf{v}) \frac{\partial \log(S_k(\mathbf{p}(\mathbf{v})))}{\partial p_i(\mathbf{v})} = -\theta + \theta = 0.$$

Using this fact, the first order condition boils down to

$$v_i + \boldsymbol{\xi}_i(\mathbf{v}) - \boldsymbol{\gamma}(\mathbf{v}) + \lambda \left(\log(S_i(\mathbf{p}^0)) - \log(S_i(\mathbf{p}(\mathbf{v})))\right) \quad = \quad 0.$$

The previous expression may be rewritten as

$$\lambda \left(\log(S_i(\mathbf{p}(\mathbf{v}))) - \log(S_i(\mathbf{p}^0))\right) = v_i + \boldsymbol{\xi}_i(\mathbf{v}) - \boldsymbol{\gamma}(\mathbf{v}).$$

Then

$$\frac{S_i(\mathbf{p}(\mathbf{v}))}{S_i(\mathbf{p}^0)} \quad = \quad e^{\frac{1}{\lambda}(v_i + \boldsymbol{\xi}_i(\mathbf{v}) - \boldsymbol{\gamma}(\mathbf{v}))}.$$

Noting that for an interior solution we must have $\boldsymbol{\xi_i(\mathbf{v})} = 0$

$$S_i(\mathbf{p}(\mathbf{v})) \quad = \quad S_i(\mathbf{p}^0) e^{\frac{1}{\lambda}(v_i - \boldsymbol{\gamma}(\mathbf{v}))}.$$

Under **C1** and **C4**, and defining $\mathbf{S} = \mathbf{H}^{-1}$ we find:

$$p_i(\mathbf{v}) \quad = \quad H_i(e^{(\mathbf{v} + \log \mathbf{S}(\mathbf{p}^0))/\lambda}) \cdot e^{\frac{-\boldsymbol{\gamma}(\mathbf{v})}{\lambda}}. \tag{29}$$

Adding up

$$\sum_{i=1}^{N} p_i(\mathbf{v}) \quad = \quad \sum_{i=1}^{N} H_i(e^{(\mathbf{v} + \log \mathbf{S}(\mathbf{p}^0))/\lambda}) \cdot e^{\frac{-\boldsymbol{\gamma}(\mathbf{v})}{\lambda}}.$$

It follows then that

$$e^{\frac{\boldsymbol{\gamma}(\mathbf{v})}{\lambda}} \quad = \quad \sum_{i=1}^{n} H_i(e^{(\mathbf{v} + \log \mathbf{S}(\mathbf{p}^0))/\lambda}).$$

Finally we get

$$p_i(\mathbf{v}) \quad = \quad \frac{H_i(e^{(\mathbf{v} + \log \mathbf{S}(\mathbf{p}^0))/\lambda})}{\sum_{j=1}^{n} H_j(e^{(\mathbf{v} + \log \mathbf{S}(\mathbf{p}^0))/\lambda})}.$$

□

Proof of Corollary ??. Matejka and McKay (2015) shows that exchangeability assumption, $p_i^0 = \frac{1}{N}$. Using **C1** in (19) the result follows at once.

Proof of Proposition 5. First, consider the RI problem (17), and for simplicity assume $\lambda = 1$. Under assumption 2, by proposition 4 we know that there exists a solution to (17). In particular,

for each value of $\mathbf{v}$, the optimal solution $\mathbf{p}(\mathbf{v})$ satisfies:

$$p_i(\mathbf{v}) = \frac{H_i(e^{\mathbf{v}+\log \mathbf{S}(\mathbf{p}^0)})}{\sum_{j=1}^N H_j(e^{\mathbf{v}+\log \mathbf{S}(\mathbf{p}^0)})} \quad \text{for } i = 1, \ldots, N.$$

Defining $\tilde{v}_i = v_i + \log S_i(\mathbf{p}^0)$ for $i = 1, \ldots, N$, and under assumption 1, it follows that $\mathbf{p}(\mathbf{v}) = \nabla W(\tilde{\mathbf{v}})$. But latter expression is just $\mathbf{p}(\mathbf{v}) = \mathbf{q}(\tilde{\mathbf{v}})$.

Now, lets consider the discrete choice model defined by Eqs. (3),(6). Assuming that assumption 1 holds, we may plug in (20) in Eq. (3) to obtain the choice probabilities:

$$\begin{aligned} q_i(\tilde{\mathbf{v}}) &= \frac{H_i\left(e^{\tilde{\mathbf{v}}}\right)}{\sum_{j=1}^N H_j\left(e^{\tilde{\mathbf{v}}}\right)} \quad \text{for } i = 1, \ldots, N, \\ &= \frac{H_i(e^{\mathbf{v}+\log \mathbf{S}(\mathbf{p}^0)})}{\sum_{j=1}^N H_j(e^{\mathbf{v}+\log \mathbf{S}(\mathbf{p}^0)})}. \end{aligned}$$

But latter expression corresponds to the solution to the rational inattention problem (17).

Thus we conclude that for all $\mathbf{v} \in \mathcal{V}$,

$$p_i(\mathbf{v}) = q_i(\tilde{\mathbf{v}}) \quad \text{for } i = 1, \ldots, N.$$

□

Proof of proposition 6.

(i) Combining Proposition 2 and Eq. (18), we get (23).
(ii) For each $\mathbf{v} \in \mathcal{V}$, let $\mathbf{q}(\tilde{\mathbf{v}})$ be the choice probability vector generated by a discrete choice model consistent with Eqs. (1)-(4). Looking at the optimization problem (17) s.t. (9) and (10) it is easy to see that it can be written as

$$\max_{\{p_i(\mathbf{v})\}_{i=1}^N} \sum_{\mathbf{v}\in\mathcal{V}} \left(\sum_{i=1}^N p_i(\mathbf{v}) \cdot (v_i + \log S_i(\mathbf{p}^0)) - \sum_{i=1}^N p_i(\mathbf{v}) \log S_i(\mathbf{p}(\mathbf{v})) \right) \mu(\mathbf{v}).$$

Then for each value $\mathbf{v} \in \mathcal{V}$, a solution to the RI program may solve following (pointwise) optimization problem:

$$\max_{p_i(\mathbf{v})\in\Delta} \left\{ \sum_{i=1}^N p_i(\mathbf{v}) \cdot (v_i + \log S_i(\mathbf{p}^0)) - \sum_{i=1}^N p_i(\mathbf{v}) \cdot \log S_i(\mathbf{p}(\mathbf{v})) \right\}.$$

Then by Proposition 2, the previous problem is equivalent to

$$\max_{p_i(\mathbf{v})\in\Delta} \left\{ \sum_{i=1}^N p_i(\mathbf{v}) \cdot (v_i + \log S_i(\mathbf{p}^0)) - W^*(\mathbf{p}) \right\}.$$

Then the conclusion follows at once.

(*iii*) Let $\mathbf{p}^*(\mathbf{v})$ be a solution to the RI program. At the optimum (and taking $\lambda = 1$), it is easy to see that the RI objective function in Eq. (17) can be written as:

$$\sum_{\mathbf{v}\in\mathcal{V}} \left(\sum_{i=1}^{N} p_i^*(\mathbf{v})(v_i + \log S_i(\mathbf{p}^0)) - \sum_{i=1}^{N} p_i^*(\mathbf{v}) \log S_i(\mathbf{p}^*(\mathbf{v})) \right) \mu(\mathbf{v}).$$

By proposition 5 we know that $\mathbf{p}^* = \mathbf{q}(\tilde{\mathbf{v}})$. This latter fact implies that the RI program can be rewritten as

$$\sum_{\mathbf{v}\in\mathcal{V}} \left(\sum_{i=1}^{N} q_i(\tilde{\mathbf{v}})\tilde{v}_i - \sum_{i=1}^{N} q_i(\tilde{\mathbf{v}}) \log S_i(\mathbf{q}(\tilde{\mathbf{v}})) \right) \mu(\mathbf{v}).$$

Now, by proposition 2 it follows that $W(\mathbf{v} + \log \mathbf{S}(\mathbf{p}^0)) = \sum_{i=1}^{N} q_i(\tilde{\mathbf{v}})\tilde{v}_i - W^*(\mathbf{q}(\tilde{\mathbf{v}}))$. Thus the optimum value of the RI program is given by:

$$\sum_{\mathbf{v}\in\mathcal{V}} W(\mathbf{v} + \log \mathbf{S}(\mathbf{p}^0))\mu(\mathbf{v}).$$

□

Proof of proposition 7.

For convenience, define $z_c(\mathbf{v}) = \exp(v_c)$. Assume, towards a contradiction, that $p_a^0 > 0$. Then

$$p_a^0 = \mathbb{E}_{\mathbf{v}} \left(\frac{H_a\left(\left\{z_c\left(\mathbf{v}\right) S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^{N}\right)}{\sum_b H_b\left(\left\{z_c\left(\mathbf{v}\right) S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^{N}\right)} \right) \tag{30}$$

$$< \mathbb{E}_{\mathbf{v}} \left(\frac{H_a\left(\left\{z_a\left(\mathbf{v}\right) S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^{N}\right)}{\sum_b H_b\left(\left\{z_a\left(\mathbf{v}\right) S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^{N}\right)} \right) \tag{31}$$

$$= \mathbb{E}_{\mathbf{v}} \left(\frac{z_a(\mathbf{v}) H_a\left(\left\{S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^{N}\right)}{z_a(\mathbf{v}) \sum_b H_b\left(\left\{S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^{N}\right)} \right) = \mathbb{E}_{\mathbf{v}} \left(\frac{p_a}{\sum_b p_b} \right) = p_a. \tag{32}$$

The penultimate equality (32) follows from the homogeneity of $\mathbf{H}$ and $\mathbf{H} = \mathbf{S}^{-1}$. The first inequality (31) follows from cyclic monotonicity, which is a property of the gradient of convex functions. (See, for instance, Rockafellar (1970, Thm. 23.5).) Since the surplus function W is convex, its gradient, corresponding to the choice probabilities $\mathbf{p}(\cdot)$ is a cyclic monotone mapping. Cyclic monotonicity implies that

$$\left\langle \mathbf{p}\left(\left\{z_a\left(\mathbf{v}\right) S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^{N}\right) - \mathbf{p}\left(\left\{z_c\left(\mathbf{v}\right) S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^{N}\right), \left\{z_a\left(\mathbf{v}\right) S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^{N} - \left\{z_c\left(\mathbf{v}\right) S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^{N} \right\rangle \geq 0.$$

All the terms on the second term on the LHS are ≤ 0, except for the a-th term, which is equal to zero. In order to satisfy the inequality, then, we must have $p_a\left(\left\{z_a\left(\mathbf{v}\right) S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^{N}\right) \geq p_i\left(\left\{z_c\left(\mathbf{v}\right) S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^{N}\right)$

with the inequality strict with positive probability. Otherwise, we would have

$$\sum_{i\neq a}\left\{p_i\left(\left\{z_a\left(\mathbf{v}\right)S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right)-p_i\left(\left\{z_c\left(\mathbf{v}\right)S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right)\right\}>0$$

which leads to

$$\begin{aligned}
&\left\langle \mathbf{p}\left(\left\{z_a\left(\mathbf{v}\right)S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right)-\mathbf{p}\left(\left\{z_c\left(\mathbf{v}\right)S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right),\left\{z_a\left(\mathbf{v}\right)S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N-\left\{z_c\left(\mathbf{v}\right)S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right\rangle\\
=&\sum_{c\neq a}\left(p_c\left(\left\{z_a\left(\mathbf{v}\right)S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right)-p_c\left(\left\{z_c\left(\mathbf{v}\right)S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right)\right)*\left(z_a\left(\mathbf{v}\right)-z_c\left(\mathbf{v}\right)\right)S_c\left(\mathbf{p}^0\right)\\
\leq&\max_{c\neq a}\left[\left(z_a\left(\mathbf{v}\right)-z_c\left(\mathbf{v}\right)\right)S_c\left(\mathbf{p}^0\right)\right]*\sum_{c\neq a}\left(p_c\left(\left\{z_a\left(\mathbf{v}\right)S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right)-p_c\left(\left\{z_c\left(\mathbf{v}\right)S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right)\right)\\
\leq&0
\end{aligned}$$

with the inequality strict with positive probability. Hence, we conclude that $p_a=0$. □